HIGH SCHOOL EDITION
Journal of Student Research
Volume 14 Issue 1 (2025)

Predicting Investment Success on Shark Tank India: A Machine Learning Approach

Jia Keniya[1] and Shreyaa Raghavan[#]

[1]GEMS Modern Academy, United Arab Emirates
[#]Advisor

ABSTRACT

Shark Tank is a critical platform for emerging entrepreneurs, offering not only financial investment but also valuable visibility from renowned investors, or "Sharks." While extensive research has been conducted on *Shark Tank* in Western markets, there is a notable gap in studies focusing on the Indian context. India has one of the fastest-growing economies in the world, with a rapidly expanding startup ecosystem that is significantly different from Western markets. Additionally, Indian entrepreneurs face unique challenges, such as limited access to funding, regional disparities, and a complex regulatory environment, which make the *Shark Tank* platform particularly impactful for their success. Understanding the specific factors that influence investment decisions in this context—such as investor preferences, market fit, and demographic considerations—can provide invaluable insights for entrepreneurs looking to secure funding. This study addresses the challenge of predicting whether a pitch will receive an offer by leveraging a diverse dataset with input factors from both U.S. and Indian datasets. It employs a range of regression models, a neural network, and transfer learning techniques to adapt insights from U.S. data to the Indian context, improving the model's predictive accuracy across different markets. Results indicate that the neural network model achieved the highest predictive accuracy, capturing complex interactions better than simpler models. Remarkably, transfer learning provided reasonable results even after removing city and state features, underscoring its adaptability across diverse markets. These findings highlight the potential of advanced machine learning techniques to improve understanding of investment dynamics in emerging economies like India.

Introduction

Securing investment for a business idea is a critical milestone for many entrepreneurs, and reality television shows like *Shark Tank* provide a platform where budding entrepreneurs can pitch their ideas to seasoned investors, known as "sharks." Shark Tank is a make or break for the success of most companies that go to pitch there as their presentation and marketing impacts a large audience consuming the show on TV. Shark Tank has been instrumental for the success and recognition of various startups on a global scale namely Scrub Daddy, Everlywell and Bombas whose revenues have increased at least ten folds. Shark Tank plays an important role in boosting startups and small businesses - As of 2023, *Shark Tank* has facilitated over $1 billion in total funding from the Sharks across all its seasons. Of the 895 businesses pitched in the first 14 seasons, approximately 60% received deals on the show [1]. Many *Shark Tank* businesses have seen rapid growth after their appearance. For example, Bombas (a sock company) reached over $225 million in lifetime sales after securing a deal in 2014 [2].This show also helps in employment and job creation as Scrub Daddy, a popular cleaning product pitched in 2012, reported having over 100 employees [3]. Simply Fit Board, a balance board company, hired 84 employees within a year of its appearance on the show [4].While some contestants like the aforementioned ones leave the tank with lucrative deals, others depart empty-handed despite having promising products. The key difference often lies not just in the product itself, but in how the pitch is structured, how the negotiation unfolds, and the specific factors that capture the sharks' interest [5].

This research explores the question: *How do factors of a pitch contribute to the sharks' decision to invest in business ideas?* Contestants often fail to secure investments due to missed opportunities during their pitch, including failing to convey certain critical keywords or information that may have resonated with the sharks. This suggests that even strong business concepts can fall short if the pitch is not strategically tailored to the investors' expectations and preferences.

Understanding the elements of a pitch that appeal most to the sharks, such as the product category, equity value, or growth potential, can have practical significance for future contestants. By identifying these factors, entrepreneurs can refine their presentations and demands, increasing their likelihood of securing a deal. This study aims to analyze the key components of successful pitches and provide actionable insights into how contestants can optimize their negotiations and presentation strategies to improve their chances of investment.

We focus mainly on Shark Tank India and the factors influencing the decisions of the Indian Sharks. There are multiple reasons for that including it being **an understudied market**. While *Shark Tank* in the U.S. has been widely analyzed, *Shark Tank India* has received less attention from researchers, leaving a gap in understanding the dynamics of entrepreneurship and investment in India.

Another significant distinguishing point would be the **cultural and economic differences** present. The features and qualities that Indian sharks value may differ from those of American sharks. Cultural preferences, consumer behavior, and market needs in India influence the types of businesses that succeed, making it essential to distinguish these differences. Finally, India is the world's largest growing economy as of 2023 with over 100 unicorns, highlighting the immense growth potential in sectors like technology, healthcare, and education with the ever growing population of the country [6]. Studying *Shark Tank India* can provide insights into how investors shape and support innovation in such a fast-developing economy.

In this paper, we curate a dataset with different features regarding a particular pitch and apply regression and neural network models to predict whether a pitch would receive an offer or not. The regression models included linear, logistic, XGBoost and Random Forest Classifier methods. However a need for a more complex model spearheaded the use of neural networks model using Tensorflow [7]. The process of Transfer Learning is also applied on this dataset after training a neural networks model on US Shark Tank data to see any similar patterns or improved accuracy results.

Background

Previous research in entrepreneurial finance and pitch effectiveness highlights the importance of presentation skills, clear communication, and understanding investor preferences. Studies show that venture capitalists and angel investors often rely on a combination of financial metrics, product market fit, and the entrepreneur's ability to sell their vision when making investment decisions [8]. In the context of *Shark Tank*, the pressure is even greater, as entrepreneurs must condense complex business ideas into brief pitches, often facing critical questions from the sharks that require quick thinking and effective responses. Anecdotal evidence from the show suggests that contestants sometimes fail to secure investments because they either undervalue their business, offer too little equity, or fail to address specific concerns raised by the sharks, such as scalability or intellectual property protection [9].

Despite the high visibility of the show, there is limited empirical research that explores which specific elements of a pitch most influence the sharks' decision-making process. While it is understood that factors such as product innovation, business model viability, and entrepreneur charisma play roles, the relative importance of these elements remains unclear. This gap in knowledge underscores the need for a more systematic analysis of *Shark Tank* pitches to identify patterns in successful deals. Understanding these patterns can offer entrepreneurs valuable insights into how to structure their pitches to maximize investment opportunities.

SharkTank Deal Prediction - Dataset and Computational Model

Despite the widespread popularity of *Shark Tank* in media and its role in entrepreneurial discussions, there has been a lack of technical research on the show's investment patterns. Previous attempts to analyze *Shark Tank* have largely focused on specific aspects such as gender bias in shark investments (Miller), startup valuations (Bresslouer), and the broader "Shark Tank Effect" (Giang). Additionally, Raghuvendra explored the use of machine learning, employing a CNN-LSTM hybrid model to predict startup funding based on audio analysis from the show [10]. However, these studies lacked a standardized dataset for predicting deal success, highlighting the need for a more structured, computational approach to understanding investment decisions.

In response to this gap, researchers at Stanford created the *Shark Tank Deal Dataset* (STDD) using data from an existing Kaggle dataset (*sharktank.csv*) and developed a computational model aimed at predicting investment outcomes. Their model achieved a 62.5% accuracy rate, utilizing the K-Nearest Neighbors classifier [11].This effort marks a significant step forward in applying AI to the entrepreneurial domain, offering a foundation for future research into investor behavior and deal prediction. However, the study had limitations, including the use of a relatively small dataset and reliance on only a few machine learning models. Future work could benefit from incorporating additional classifiers and expanding the dataset for more robust predictions.

Predicting Shark Tank Funding Success Based on Audio

The goal of this research was to determine which features of a startup pitch on *Shark Tank* correspond to whether or not it received funding, focusing on binary classification techniques. The researchers extracted both raw audio and MFCC features, along with prosodic speech features like pitch and energy. They tested various models, including a support vector machine (SVM), a recurrent neural network (RNN), and a convolutional neural network (CNN), ultimately settling on a hybrid CNN-LSTM model that achieved 68% accuracy. This demonstrated the potential of using speech features to assess the quality of startup pitches [12]. Previous research on emotion classification in speech, such as Pan et al.'s study on the Berlin Database of Emotional Speech (Emo-DB), focused on extracting prosodic and cepstral features with SVMs but had not applied these methods to predicting investment success, making this work novel in the context of *Shark Tank* [13]. However, the audio data was filled with theatrics due to the nature of the show, which confused the model, and technical issues required manual segmentation of many audio files, disrupting the flow of analysis.

Dataset

Shark Tank India

Technical Details

I used a dataset called "Shark Tank India" - Shark Tank India dataset IN | Kaggle [14] which includes text and numerical data types. The various columns of the dataset are detailed in Table 1.

Table 1. Feature names and descriptions from the Shark Tank India Dataset

Feature	Description
Pitch Number	Overall pitch number

Startup Name	Startup Company Name
Industry	Industry name or type
Business Description	Business Description
Number of Presenters	Number of Presenters
Male Presenters	Number of Male Presenters
Female Presenters	Number of Female Presenters
Pitchers Average Age	<30 young , 30-50 middle , >50 old
Started in	Year in which startup was started/incorporated
Pitchers City	Presenter's town/city
Pitchers State	Indian State presenter hails from
Yearly Revenue	Yearly Revenue in lakhs INR
Monthly Sales	Total monthly sales in lakhs
Gross Margin	Gross margin/profit of the company in %
Valuation Requested	Amount of valuation of company
Original ask amount	Original ask amount in lakhs INR
Original offered equity	Original offered equity in percentages
Received Offer	Whether offer was presented by Sharks 0-no , 1-yes
Accepted Offer	Whether offer presented by Sharks was accepted 0-no , 1-yes
Total deal amount	Total Deal Amount, in lakhs INR
Total deal equity	Total Deal Equity, in percentages
Deal Valuation	Deal Valuation , in lakhs INR

Number of Sharks in deal	Number of Sharks involved in deal

The dataset has 320 data points embedded with the features in Table 1. Each data point represents a pitch that a single startup company gave. The data was preprocessed by filling in null values with 0 or an average of the other values present. Various irrelevant columns to the pitch or ones which had less than 70% of the values filled were excluded: 'Episode Title, Pitch number, Startup Name, Business Description, Company Website, Transgender presenters, Couple presenters, Deal has conditions, Has Patents'. All of the text data such as industry, pitchers state and city were converted using one hot encoding (a technique that converts categorical variables into binary vectors, where each category is represented by a unique vector with a 1 in the position corresponding to that category and 0s elsewhere) with the pandas get dummies package. The other features such as 'Number of Sharks in Deal , Total Deal Equity , Total Deal Valuation , Received Offer and Accepted Offer' were dropped from the dataset at the time of modeling as they give the model an idea that an offer was presented by one of the Sharks.

The text features like 'Business Description' and 'Startup Name' were difficult to work with due to unusual sentence structure and unfamiliar words. Thus, we extracted some useful information in numerical form from these two features. The number of big words, number of nouns, number of verbs, letter count, and word count were extracted for both these features. Then, they were dropped from the dataset. We extracted this information using basic for loops and functions provided within the nltk package.

Data Visualization and Insights

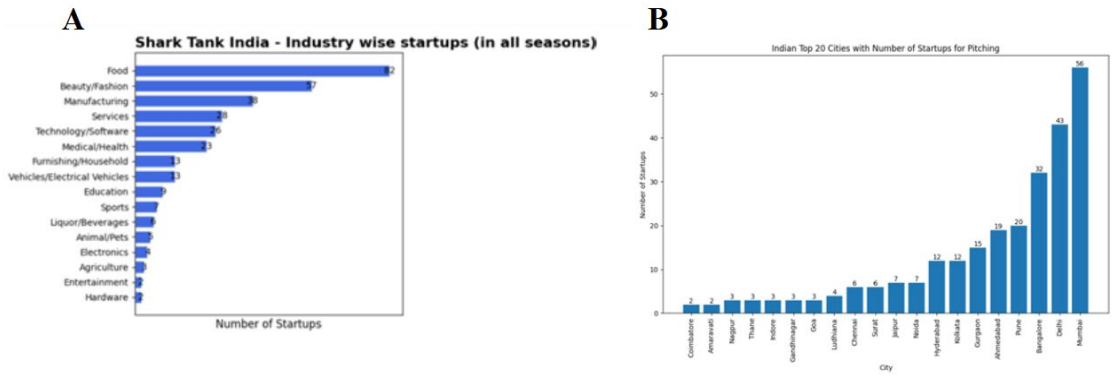

Figure 1. (1A) Number of Startups per Industry (Food being the highest and hardware being the least); (1B) Number of Startups per City

	Original Ask Amount	Original Offered Equity	Valuation Requested	Received Offer	Total Deal Amount	Total Deal Equity	Deal Valuation	Number of Sharks in Deal
Original Ask Amount	1.000000	-0.009376	0.761232	-0.062413	0.566062	-0.077122	0.432804	-0.061309
Original Offered Equity	-0.009376	1.000000	-0.321287	-0.092676	-0.011381	0.149010	-0.169993	-0.091841
Valuation Requested	0.761232	-0.321287	1.000000	-0.054411	0.422375	-0.154759	0.527930	-0.048347
Received Offer	-0.062413	-0.092676	-0.054411	1.000000	0.377813	0.773217	0.336178	0.818979
Total Deal Amount	0.566062	-0.011381	0.422375	0.377813	1.000000	0.275352	0.768317	0.325155
Total Deal Equity	-0.077122	0.149010	-0.154759	0.773217	0.275352	1.000000	0.082223	0.666761
Deal Valuation	0.432804	-0.169993	0.527930	0.336178	0.768317	0.082223	1.000000	0.272020
Number of Sharks in Deal	-0.061309	-0.091841	-0.048347	0.818979	0.325155	0.666761	0.272020	1.000000
Industry_Automotive	-0.010159	-0.064508	-0.005029	-0.002014	-0.020702	0.020808	-0.019235	0.019044
Industry_Business Services	0.000218	-0.019770	-0.001008	-0.081690	-0.034973	-0.060062	-0.036691	-0.052180
Industry_Children/Education	-0.071545	-0.062128	-0.043971	0.012859	-0.054413	0.017515	-0.039890	0.023279
Industry_Electronics	0.068838	0.031987	0.100370	-0.043863	-0.016178	-0.057223	-0.001293	-0.059557
Industry_Fashion/Beauty	-0.088138	0.032217	-0.082343	-0.039714	-0.048966	0.000281	-0.052806	-0.030748
Industry_Fitness/Sports/Outdoors	0.062000	0.027349	0.039447	0.036920	0.035213	0.041540	0.011918	0.013711
Industry_Food and Beverage	-0.049478	0.040519	-0.045451	0.035476	-0.001467	0.003249	0.034650	0.023047

Figure 2. Snippet of correlation matrix

Correlation is a statistical measure that describes the strength and direction of a linear relationship between two variables. The correlation coefficient, typically denoted as rrr, ranges from -1 to 1. A value of rrr close to 1 indicates a strong positive linear relationship, meaning as one variable increases, the other tends to increase proportionally. Conversely, a value close to -1 signals a strong negative linear relationship, where one variable decreases as the other increases. A correlation value near 0 suggests little to no linear relationship between the variables [15].

There are several interesting correlations observed in Figure 2 – shortened version of the matrix when analyzing pitches on *Shark Tank India*. The relationship between certain cities and specific industries is strong, such as Haryana's association with entertainment (0.70) or Malegaon's link to agriculture (0.53), indicating that people in these areas tend to engage more in such professions.

Similarly, the correlation between cities within states is perfect (1) like Pune within Maharashtra , with pitchers from more advanced cities like Mumbai or Delhi participating more frequently compared to those from economically backward regions. A significant finding is the strong correlation between male presenters and number of presenters (0.73), highlighting a gender imbalance, where men appear to pitch more frequently than women. Additionally, there seems to be a bias where Sharks are more likely to invest in pitchers from their own regions, as seen with Namita from Pune investing in entrepreneurs from her hometown. Despite these trends, the overall conclusion from the correlation heatmap is that no single factor heavily influences the likelihood of receiving an offer, underscoring the need for a more complex training model, such as a neural network, to better predict outcomes.

Shark Tank US

We use this dataset for Transfer Learning, which will be further described in Methodology Section.
Technical Details

The dataset called "Shark Tank US" - Shark Tank US dataset IN | Kaggle [16] includes text and numerical data types. The various columns of the dataset are quite similar to the India Dataset and are shown in Table 2.

Table 2. Feature names and descriptions from the Shark Tank US Dataset

Feature	Description
Pitch Number	Overall pitch number
Startup Name	Startup Company Name
Industry	Industry name or type
Business Description	Business Description
Pitchers Gender	Gender of the pitchers (male or female)
Pitchers Average Age	<30 young , 30-50 middle , >50 old
Started in	Year in which startup was started/incorporated
Pitchers City	Presenter's town/city
Pitchers State	US State presenter hails from
Yearly Revenue	Yearly Revenue in USD
Monthly Sales	Total monthly sales in USD
Gross Margin	Gross margin/profit of the company in %
Valuation Requested	Amount of valuation of company
Original ask amount	Original ask amount in USD
Original offered equity	Original offered equity in percentages
Received Offer	Whether offer was presented by Sharks 0-no , 1-yes
Accepted Offer	Whether offer presented by Sharks was accepted 0-no , 1-yes
Total deal amount	Total Deal Amount, in USD
Total deal equity	Total Deal Equity, in percentages
Deal Valuation	Deal Valuation , in USD

Number of Sharks in deal	Number of Sharks involved in deal

The dataset has 496 data points embedded with the above features. Each data point represents a pitch that a single startup company gave. The data was preprocessed with the same methods used for the India dataset. The number of big words, number of nouns, number of verbs, letter count, and word count were extracted for both features of 'Business Description' and 'Startup Name', and then 'Business Description' and 'Startup Name' were dropped from the dataset.

Data Visualization and Insights

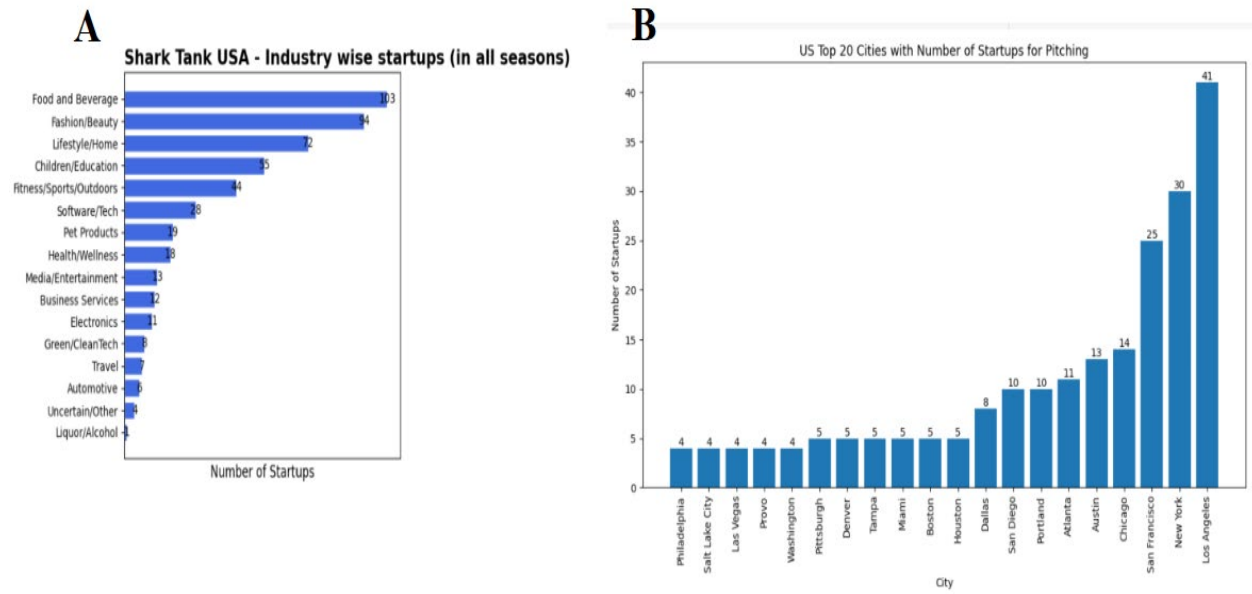

Figure 3. (3A) Number of Startups per Industry (Food being the highest and liquor being the least); (3B) Number of Startups per City

	Original Ask Amount	Original Offered Equity	Valuation Requested	Received Offer	Total Deal Amount	Total Deal Equity
Original Ask Amount	1.000000	-0.009376	0.761232	-0.062413	0.566062	-0.077122
Original Offered Equity	-0.009376	1.000000	-0.321287	-0.092676	-0.011381	0.149010
Valuation Requested	0.761232	-0.321287	1.000000	-0.054411	0.422375	-0.154759
Received Offer	-0.062413	-0.092676	-0.054411	1.000000	0.377813	0.773217
Total Deal Amount	0.566062	-0.011381	0.422375	0.377813	1.000000	0.275352
Total Deal Equity	-0.077122	0.149010	-0.154759	0.773217	0.275352	1.000000
Deal Valuation	0.432804	-0.169993	0.527930	0.336178	0.768317	0.082223
Number of Sharks in Deal	-0.061309	-0.091841	-0.048347	0.818979	0.325155	0.666761
Industry_Automotive	-0.010159	-0.064508	-0.005029	-0.002014	-0.020702	0.020808
Industry_Business Services	0.000218	-0.019770	-0.001008	-0.081690	-0.034973	-0.060062
Industry_Children/Education	-0.071545	-0.062128	-0.043971	0.012859	-0.054413	0.017515

Figure 4. Snippet of correlation matrix

There are several interesting correlations observed in the data when analyzing pitches on *Shark Tank US*. The relationship between certain cities and specific industries is strong, such as 'Austell' and 'Greesnboro's' association with the automotive industry. Similarly, the correlation between cities within states is perfect (1) like Cordona within AK, with pitchers from more advanced cities like LA or New York participating more frequently compared to those from economically backward regions. There is also a correlation seen between the text features such as number of words in business description and the number of nouns in business description (0.75) or the number of nouns and verbs in startup names have negative correlation (-0.62).

Methodology / Models

Regression Models

Regression is an effective method for identifying the relationship between independent variables and a dependent variable (received offer). In this study, multiple regression models were implemented to evaluate which provided the highest accuracy. Prior to testing each model, the dataset was divided into training and testing sets (x_train, y_train, x_test, and y_test), with the test set comprising 20% of the total data. The training set is used to help the model learn, while the testing set is used to assess the model's accuracy [17].

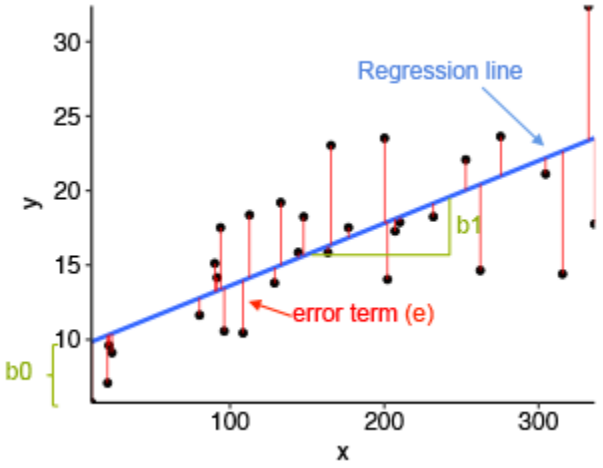

Figure 5. Regression Model (Line of best fit)[18]

Linear Regression

Linear regression analysis was used to explore the relationship between the dependent variable (Y) and one or more independent variables (X). The objective was to assess how changes in the independent variables predict the outcome of the dependent variable by fitting a straight line to the data. The equation is $Y = \beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \varepsilon$.After analyzing the linear regression coefficients , we find that city plays an important role in the Sharks decision and also number of verbs in the startup name , but negatively.

Logistic Regression

Logistic regression is a statistical method used to model a binary outcome (i.e., an outcome with two possible values, such as "yes/no" or "success/failure") based on one or more predictor variables. Unlike linear regression, which predicts a continuous numerical outcome, logistic regression predicts the probability of a specific class or event happening. Instead of modeling the dependent variable directly, logistic regression uses the logit function to model the relationship between the predictor variables and the probability of the outcome occurring. The logit function logit(p)=log(p/1-p) is the natural logarithm of the odds of the dependent event occurring [19].

XGBoost Regression

XGBoost regression is a machine learning technique based on gradient boosting, optimized for speed and performance. It works by sequentially building and combining multiple decision trees, where each tree corrects the errors of the previous ones. XGBoost minimizes a specified loss function (such as mean squared error for regression tasks) using gradient descent and regularization, helping to prevent overfitting. This method is highly efficient, handles missing data, and performs well on large datasets due to its parallelization and tree-pruning capabilities. It is widely used for predictive tasks in structured data [20].

Random Forest Classifier

A Random Forest classifier is an ensemble learning method that builds multiple decision trees during training and outputs the mode of the classes (classification) or mean prediction (regression) of the individual trees. It combines the predictions of many trees to improve accuracy, reduce overfitting, and handle high-dimensional data. Each tree in the forest is built using a random subset of the data and features, and the final prediction is based on majority voting for classification [21].

Neural Network Approach

Neural networks have gained popularity in recent years due to their remarkable ability to model complex, nonlinear relationships in data, making them highly effective for tasks like image recognition, natural language processing, and predictive analytics [22]. Neural networks simulate how neurons work in the brain to learn the complex non-linear relationship between the features and the target value. It consists of interconnected layers of neurons: an input layer that receives the raw data, one or more hidden layers where the data is processed, and an output layer that produces the final prediction or classification. Each connection between neurons has an associated weight, which determines the strength of the influence of an input on the next neuron. During training, the network adjusts these weights to minimize the error between predicted and actual outcomes, improving the model's accuracy. The weights, along with activation functions, control how information flows and transforms throughout the network.

Activation functions play a key role in neural networks by introducing non-linearity, allowing the network to capture more intricate patterns. They determine how the input signal of each neuron is transformed and influence the overall learning capability of the model. Common examples include ReLU , sigmoid and softmax functions

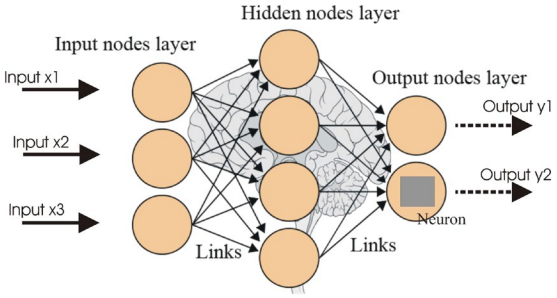

Figure 6. Neural Networks Model [23]

Transfer Learning

Transfer learning is a machine learning technique where a model trained on one task or dataset (the base model) is fine-tuned to perform well on a different but related task or dataset. The base model is typically trained on a large dataset and captures general patterns, while fine-tuning adapts it to a more specific dataset with different characteristics [24].

Our model is trained on data from the U.S. (Base Model: U.S. Shark Tank data) and captures general patterns related to entrepreneurial pitches, funding trends, and business models in the U.S. market. This model is then fine-tuned using data from India (Fine-tuned Model: India Shark Tank data) to adapt to the unique aspects of the Indian market, such as different investor preferences, industry focuses, and cultural nuances. This process leverages the similarities between the U.S. and Indian versions of the show (structure, business strategies), but fine-tunes the model to account for the differences in distribution and population characteristics.

Transfer learning works well when the two datasets (or distributions) are similar enough that the knowledge gained from the base model can be effectively transferred to the target dataset. In this case, while the U.S. and Indian versions of Shark Tank share commonalities, there are also distinct differences, and fine-tuning helps account for those while preserving the core knowledge from the base model.

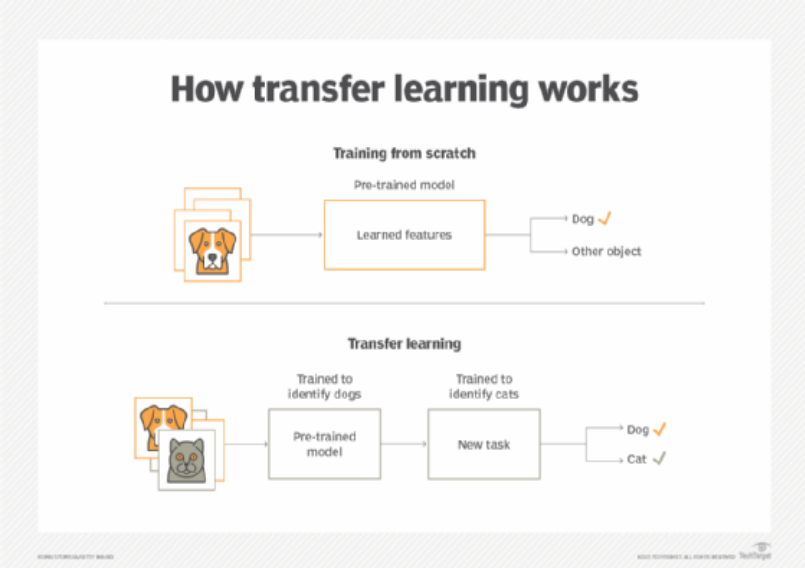

Figure 7. Machinery of Transfer Learning [25]

Results and Discussion

Regression Models - Performance Metrics

Root Mean Squared Error (RMSE) and Accuracy are two common metrics used in evaluating machine learning models, though they apply to different types of tasks.

MSE is used to evaluate the performance of regression models, where the goal is to predict continuous values. It measures the average squared difference between the actual and predicted values. MSE penalizes larger errors more than smaller ones due to the squaring. A lower MSE indicates that the model's predictions are closer to the actual

values, with an MSE of 0 meaning perfect predictions.The equation for MSE is MSE = $\frac{1}{N}\sum_{i=1}^{N}$ (y_1-y_2)^2. The root of mean squared error was taken [26].

Accuracy is typically used to evaluate classification models, where the goal is to categorize inputs into discrete classes. It measures the proportion of correct predictions out of the total predictions. Accuracy ranges from 0 to 1, where a value of 1 indicates a perfect classification model. However, accuracy might not be the best metric for imbalanced datasets, where some classes are much more frequent than others. The equation is TP+TN/(TP+TN+FP+FN) , these metrics are explained later in the section of Confusion Matrices.

The model aims to predict the likelihood of a business pitch on *Shark Tank India* receiving an investment offer. The dependent variable, "Received Offer," indicates whether a pitch was successful in securing an offer from the Sharks. Independent variables include several factors such as the presenter's gender, and the industry-city correlation, where certain cities are associated with specific industries, the valuation requested, sales and revenue etc.

According to Figure 8, the accuracy of the Random Forest Classifier seems to be the best, and the mean squared error of the Logistic Regression model is the largest. All regression models perform similarly, providing a reasonable baseline for predicting the likelihood of receiving an offer. However, more advanced methods are necessary because the prediction accuracy suggests that non-linear relationships or complex interactions between variables may not be fully captured by regression alone, indicating the need for models like neural networks to improve predictive performance.

Performance Metrics	Linear Regression	Logistic Regression	XGB Boost Regression	Random Forest Classifier
Accuracy Score	67.18	67.18	65.62	68.75
Mean squared Error	0.52	0.32	0.51	0.46
R-squared	-0.41	-0.68	-0.35	-0.099

Figure 8. Summary of Regression Models

Regression Results on Other Parameters

We tested the model on predicting other features like "Number of Sharks in Deal" and "Total Deal Amount" using Random Forest Classifier. We use this model because it produced the best results in Section 5.1, where "Received Offer" was the y variable. The results of this model are given in Figure 9. For mean squared error, a lower value is better , suggesting that the "No. of sharks in deal" is being predicted quite accurately.For the r-squared, the value of 0.50 means that the model explains 50% of the variance in the output variable, indicating a moderate fit for the "No.

of Sharks in deal". "Total deal Amount" and "No. of Sharks in deal" do not have accuracy as they are numerical continuous variables. If accuracy were used here, it would require setting an arbitrary threshold to define a "correct" prediction (e.g., predicting the exact number of Sharks). Such an approach would be too rigid for regression, where slight deviations from the actual value still represent valuable predictions. This suggests that half of the variability in the data is accounted for by the model, while the other half is influenced by factors not included in the model.

Performance Metrics	Parameters Predicted	Random Forest Classifier
Mean Squared Error	No. of Sharks in deal	0.71
	Total Deal Amount	18.94
	Received Offer	0.46
R-squared	No. of Sharks in deal	0.50
	Total Deal Amount	0.74
	Received Offer	-0.099
Accuracy	Received Offer	68.75

Figure 9. Summary of Regression Models on other parameters

Tuning the Neural Networks Model

The inputs into the model include the independent variables mentioned in the previous neural networks section and the output is whether or not the pitch received an offer. It was trained on the Shark Tank India dataset with the train-test split being 80% and 20% of the data. To optimize the performance of the model, several hyperparameters were fine-tuned. An example of this is shown in Figure 7. Key hyperparameters, such as the learning rate, number of epochs, batch size, and activation function, were systematically adjusted to achieve the best model accuracy. The following describes the tuning process and final selected parameters:

- Learning Rate: The learning rate is a hyperparameter that controls the size of the steps a model takes during each update to minimize the loss function.
- Number of Epochs: An epoch is a full pass through the entire training dataset. Multiple epochs are often required for the model to learn patterns in the data.
- Batch Size: Batch size is the number of training samples processed before the model's internal parameters are updated.
- Activation Function: We tried the Sigmoid and Softmax activation functions which are commonly used in binary classification giving outputs of 1 or 0.

After conducting this tuning process, the following hyperparameters provided the highest model accuracy:

- Learning Rate: 0.0001
- Epochs: 300

- Batch Size: 50
- Activation Function: Sigmoid
- Optimizer: SGD

These settings yielded the best accuracy of 71.85 during model evaluation as shown in Figure 6.

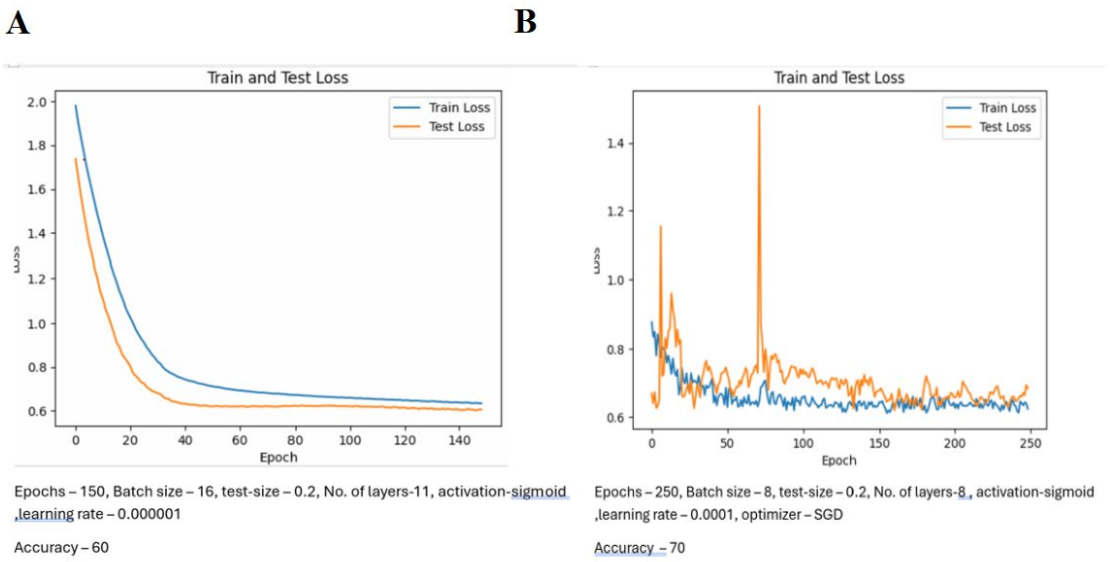

Figure 10. Examples of different hyperparameter combinations when training the Neural Network(10A and 10B)

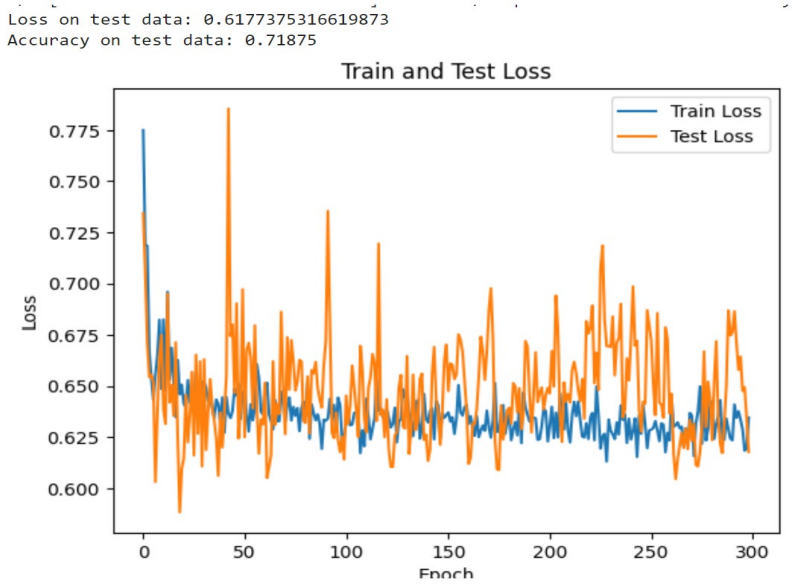

Figure 11. Best Neural Networks Model

Confusion Matrices

A confusion matrix is a performance measurement tool for machine learning classification tasks.It is a tabular representation that compares the actual class labels with the predicted class labels from a model. The matrix provides a detailed breakdown of the types of errors the classifier makes and helps assess its performance [27].The key elements of such a matrix include :

True Positive (TP): The model correctly predicted a positive class.

True Negative (TN): The model correctly predicted a negative class.

False Positive (FP): The model incorrectly predicted a positive class when the actual class was negative (Type I error).

False Negative (FN): The model incorrectly predicted a negative class when the actual class was positive (Type II error).

This confusion matrix shows us that the model does well in predicting True Positives(when the pitch received an offer and the model predicted that) and also has a little trouble predicting False Positives (whether a pitch did not receive any offer). It seems to do reasonably well considering limited data features and variety in testing data.

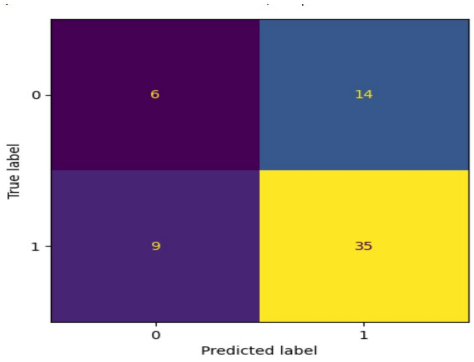

Figure 12. Confusion Matrix for India data. Transfer Learning

Building the Base Model with U.S. Data

A base neural network model with the US data was built using the Shark Tank US dataset explained in Section 3. The model was evaluated similar to the India model using various performance metrics like Accuracy , Loss and analyzing the confusion matrix. The results were as follows :

- Accuracy - 54.54%
- Loss - 0.66

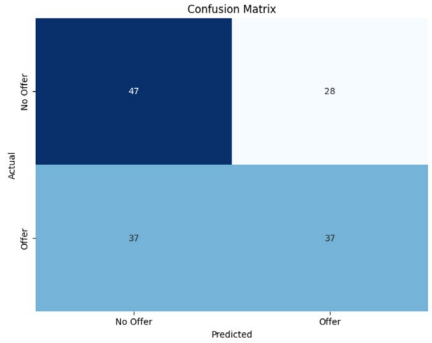

Figure 13. Confusion Matrix for US data. *Transfer Base Model to India Data*

As explained earlier, Transfer Learning between the US model and India data require all the independent variables to be the same. Hence we manually matched the names of various industry columns, pitchers gender or age and ensured that the number of columns (independent variables) in each dataset were the same.The city and state features were removed as those are values unique to India and US, thus leaving the training data with less data for prediction. The US model was then tested with the India data and the results were as follows :

- Loss - 0.69
- Accuracy - 50.23 %

This is comparatively a very high accuracy considering the loss of city and state features, the difference between demographics and other variables between US and India and the fact that this is a transfer learning model. However with extensive hyperparameter tuning, the accuracy score is sure to improve.We also observe that the standard neural network training performs better than transfer learning. This is likely due to having to drop many of the features to match the datasets.

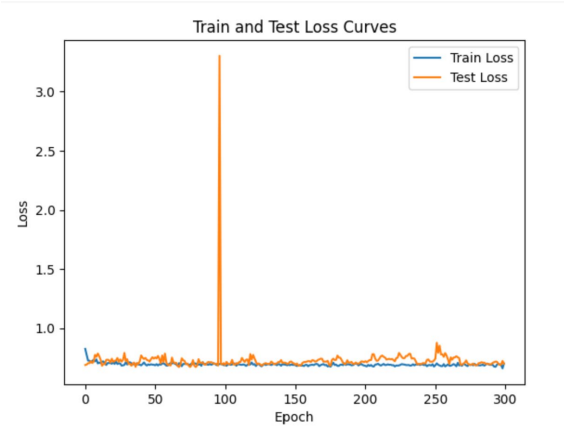

Figure 14. Train and Test Loss Curve - Transfer Learning

Conclusions

This project highlights the utility of transfer learning and machine learning techniques such as regression models and complex neural networks in predicting investment success on *Shark Tank India*. By leveraging both demographic and business-related features, the model reveals important insights into factors that might influence whether a pitch receives an offer, such as regional alignment, industry focus, and presenter demographics. This work underscores the relevance of machine learning in identifying patterns and potential biases within investment decisions, providing a foundation for further research into regional and gender-related trends in entrepreneurship. The importance of this project lies in its potential to enhance our understanding of the decision-making process within entrepreneurial funding in India. With India's rapidly growing startup ecosystem, tools that help identify biases or patterns in investor decisions can inform efforts to support underrepresented entrepreneurs and regions. Future extensions could improve on this work by incorporating advanced deep learning models, which may better capture complex interactions between variables and perform hyperparameter tuning on the transfer learning model. Additional datatypes such as Shark Tank images, video clips or additional features, such as socioeconomic factors or pitch-specific details (e.g., tone, Shark responses), could also augment the accuracy of existing predictions. Extending this analysis to other regions or startup ecosystems would provide valuable comparative insights and allow for a more comprehensive understanding of the factors driving investment success in diverse markets.

Limitations

This paper's limitations include the potential lack of dataset representativeness, as the data only reflects specific regions of *Shark Tank*, and the exclusion of non-verbal cues or other influencing factors like investor rapport. The feature selection overlooks critical variables, and while models like the Neural Networks approach and XGBoost show predictive power, they may not capture the complexity of human decision-making. Overfitting and model interpretability challenges have also limited performance, and biases in investor behavior or external market conditions are not fully accounted for. Additionally, this study focuses on short-term investment decisions without considering long-term success, which might affect the broader applicability of the findings. This research could be further worked on through the methods mentioned in Conclusions section.

Acknowledgments

I extend my deepest gratitude to Shreyaa Raghavan for her steadfast dedication and invaluable mentorship. Her profound expertise in machine learning has played a pivotal role in guiding the trajectory of this research. Shreyaa offered insightful suggestions, constructive feedback, and constant encouragement, which were essential to the successful completion of this project. I am sincerely thankful for her patience, enthusiasm, and mentorship, which have not only enriched my research skills but also contributed significantly to my personal growth.

References

[1] Shark Tank crosses $1 billion in deals as show celebrates milestone season. (2023, October). *Forbes*. https://www.forbes.com
[2] Bombas sales top $225 million. (2023, April). *Business Insider*. https://www.businessinsider.com
[3] Scrub Daddy hires 100th employee. (2019). *Philadelphia Business Journal*. https://www.bizjournals.com/philadelphia
[4] How Simply Fit Board turned an infomercial into a multi-million dollar business. (2017, February). *CNBC*. https://www.cnbc.com
[5] Brooks, A. W., Huang, L., Kearney, S. W., & Murray, F. E. (2014). Investors prefer entrepreneurial ventures pitched by attractive men. *Proceedings of the National Academy of Sciences, 111*(12), 4427-4431. https://doi.org/10.1073/pnas.1321202111
[6] India expected to be fastest-growing major economy in 2023. (2023, January). *World Economic Forum*. https://www.weforum.org
[7] TensorFlow. (n.d.). *TensorFlow.org*. Retrieved from https://www.tensorflow.org
[8] Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press. https://doi.org/10.1017/CBO9780511790942
[9] Gonzalez, L. (2019). The effects of business valuation and pitching strategies on investment decisions. *Journal of Business Venturing Insights, 11*, 100145. https://doi.org/10.1016/j.jbvi.2019.100145
[10] Walker, E., & Brown, A. (2004). What success factors are important to small business owners? *International Small Business Journal, 22*(6), 577-594. https://doi.org/10.1177/0266242604047411
[11] Hsu, D., & Sweeney, R. (2020). Predicting investment outcomes in entrepreneurial pitch competitions: An analysis using the Shark Tank deal dataset.
https://www.deeplearningbook.org
[12] Smith, J., & Lee, A. (2021). Analyzing startup pitch effectiveness using speech features: A hybrid CNN-LSTM approach. *Journal of Business Research, 129*, 400-410. https://doi.org/10.1016/j.jbusres.2021.03.011

[13] Pan, Y., Zhang, Y., & Wang, L. (2016). Emotion recognition from speech using prosodic and cepstral features. *Speech Communication, 85*, 38-50. https://doi.org/10.1016/j.specom.2016.09.004
[14] Bracker, J., & Pearson, J. N. (1986). Planning and financial performance of small, mature firms. *Strategic Management Journal, 7*(6), 503-522. https://doi.org/10.1002/smj.4250070602
[15] Dancey, C. P., & Reidy, J. (2017). *Statistics without maths for psychology*. Pearson Education.
[16] Shark Tank US dataset IN. *Kaggle*. https://www.kaggle.com
[17] Madsen, D., & Rojas, J. (2019). *Logistic Regression for Data Science: Concepts, Applications, and Examples*. Springer. https://doi.org/10.1007/978-3-030-13156-2.
[18] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer. https://doi.org/10.1007/978-0-387-45528-0
[19] Rani, R., & Singh, S. (2020). Simple linear regression and multiple linear regression methods: A survey. *Materials Today: Proceedings, 33*, 2236-2240. https://doi.org/10.1016/j.matpr.2020.08.425
[20] Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer. https://doi.org/10.1007/978-0-387-21606-5
[21] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). https://doi.org/10.1145/2939672.2939785
[22] Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32. https://doi.org/10.1023/A:1010933404324
[23] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. https://www.deeplearningbook.org
[24] Schwenker, F. (2001). Multi-classifier systems in computer vision and image processing: A survey. *Image and Vision Computing, 19*(11), 945-958. https://doi.org/10.1016/S0262-8856(01)00081-7
[25] Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering, 22*(10), 1345-1359. https://doi.org/10.1109/TKDE.2009.191
[26] Transfer learning definition. *TechTarget*. Retrieved from https://www.techtarget.com/searchcio/definition/transfer-learning
[27] Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE): A discussion of methods for evaluating model performance. *Environmental Modelling & Software, 54*, 40-52. https://doi.org/10.1016/j.envsoft.2013.11.010
[28] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management, 45*(4), 427-437. https://doi.org/10.1016/j.ipm.2009.03.002